PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development of a computable phenotype to identify a transgender sample for health research purposes: A feasibility study in a large linked provincial healthcare administrative cohort in British Columbia, Canada
AUTHORS	Rich, Ashleigh; Poteat, Tonia; Koehoorn, Mieke; Li, Jenny; Ye, Monica; Sereda, Paul; Salway, Travis; Hogg, R

VERSION 1 – REVIEW

REVIEWER	Honghan Wu UCL, UK
REVIEW RETURNED	05-Jul-2020

GENERAL COMMENTS	The paper describes a reproducible computable phenotype (CP) algorithm for identifying transgender service users from an administrative cohort in Canada. Reusable and reproducible phenotype computation is a very important area in health data research for supporting/realising reproducible studies and particularly needs community efforts. Applications of CPs in transgender subpopulation are substantially understudied and I believe this paper fits very well in filling the void. However, I think some further improvements are necessary to maximise the value of this work to the community. The performances of the phenotype algorithm are not ideal - sensitivity: <0.28 in all cases and the best PPV is 0.86 with a sensitivity around 0.09. While the poor-moderate performance is not an issue per se, it does raise the question whether the 'gold standard' is really good enough. In particular, considering ≥ 1 transgender ICD- ever has much lower PPV compared to ≥ 1 transgender ICD- recent, it seems "provider-reported transgender status" has a better coverage for recent cases. The sensitivity is particularly low in all algorithms. However, the authors did not conduct analytics to show why so. To answer/confirm this questions/suspicions, in-depth analytics by human experts are needed to look at both false positive and false negative cases. Such an effort would also generate better 'gold standard' on this cohort for transgender phenotyping algorithms. The authors did mention a related work of using natural language processing (NLP) in transgender phenotype computations. However, as significant amount of data is hidden in free-text records, this work would benefit from more discussions on this topic [1-3]: (1) whether better signals could be surfaced from free text to support better gold standard; (2) whether/how NLP and machine learning can be combined with their coded data to achieve better performances.

	It would be beneficial to link the work to the UK-wide reusable phenotype effort led by Health Data Research UK: https://portal.caliberresearch.org/. 1.Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e231. doi:10.1136/amiajnl-2013-001926 2.Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e231. doi:10.1136/amiajnl-2013-001926 11.Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350(apr24 11):h1885-h1885. doi:10.1136/bmj.h1885 3. Wu H, Hodgson K, Dyson S, et al. Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach. JMIR Medical Informatics 2019;7:e14782. doi:10.2196/14782

REVIEWER	Mei-Sing Ong Harvard Medical School & Harvard Pilgrim Health Care Institute
REVIEW RETURNED	10-Sep-2020

GENERAL COMMENTS	Ashleigh et al. developed computable phenotypes for identifying transgender patients from EHR data. This is an important contribution, given the lack of research on this population. The derived computable phenotypes achieved high specificity but poor sensitivity, which dampens my enthusiasm for the manuscript. There may be opportunities to enhance the algorithms by incorporating other diagnoses and there are also some methodological issues that need be address (please see comments below). Introduction: • The paragraph under “Methodological limitations in transgender health research” is unrelated to the subject under study. • Page 4: line 49 – the statement “computable phenotypes (CPs), also called natural language processing algorithms or case algorithms” is incorrect. Computable phenotype and natural language processing (NLP) are not synonymous. NLP is a technique for text mining, which can be applied to develop computable phenotypes. This current study did not use NLP. “Case ascertainment algorithms” may be a more appropriate term than “case algorithms”. • Page 4: line 51 – the statement “a computable phenotype is a clinical feature … that can be determined directly from EHR…” is not accurate. Rather, a computable phenotype is an algorithm for identifying clinical features using EHR data. Methods: • Since physician-assigned transgender status (i.e. the gold standard used for model validation) was available only for HIV

	patients, how did the authors validate the transgender status of non-HIV patients? • What was the mean follow-up time of each patient? The study did not appear to impose a minimum duration of follow-up in the selection criteria. Duration of follow-up can affect whether or not there was sufficient data for case ascertainment. • Page 6 (lines 39-42): "To assess face validity and utility of diagnosis and prescription data over time in CP development, concordance analyses evaluated the presence of at least one included diagnosis and prescription during the COAST study follow-up period with the presence of at least one included diagnosis and prescription in the last study year." It's unclear what was being evaluated here? Did the authors intend to evaluate whether or not diagnosis and prescription data were consistent over time? For what purpose? Please clarify. • The authors stated that concordance analysis was performed, but did not actually quantify the degree of concordance. The Kappa statistic can be a useful measure here. • Page 7 (lines 18-20): "to further assess face validity of the transgender CP for future health research, descriptive statistics were calculated …" It was not clear to me why this step was done until I read the study results. Please clarify here that a comparison was made between the study cohort vs the general population. • The authors mentioned that androgen blocker and sex hormone can be prescribed for the treatment of other conditions (e.g. hypertension, menopausal symptoms). Have the authors considered including these variables into their computable phenotypes as a way of improving the performance? Results: • In Table 1, "Unspecified endocrine disorder use" should be "unspecified endocrine disorder". • Page 8, line 4: should reference to Table 1 be Table 2 instead? Discussion: • Page 9, lines 32-37: "Though a relatively small proportion of the "true" transgender sample was identified in this study, the impact on future analyses comparing health outcomes for transgender and cisgender group is likely negligible…" I would disagree with this statement. The current algorithm missed a large proportion of transgender patients and individuals who received less healthcare services were probably more likely to be left out. And so studies that apply the derived computable phenotype to identify transgender patients may not be able to detect these disparities.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:
Reviewer: 1

Reviewer Name: Honghan Wu
Institution and Country: UCL, UK
Please state any competing interests or state 'None declared': None declared

The paper describes a reproducible computable phenotype (CP) algorithm for identifying transgender service users from an administrative cohort in Canada. Reusable and reproducible phenotype

computation is a very important area in health data research for supporting/realising reproducible studies and particularly needs community efforts. Applications of CPs in transgender subpopulation are substantially understudied and I believe this paper fits very well in filling the void. However, I think some further improvements are necessary to maximise the value of this work to the community.

Thank you, we appreciate the review.

The performances of the phenotype algorithm are not ideal - sensitivity: <0.28 in all cases and the best PPV is 0.86 with a sensitivity around 0.09. While the poor-moderate performance is not an issue per se, it does raise the question whether the 'gold standard' is really good enough. In particular, considering ≥ 1 transgender ICD- ever has much lower PPV compared to ≥ 1 transgender ICD- recent, it seems "provider-reported transgender status" has a better coverage for recent cases. The sensitivity is particularly low in all algorithms. However, the authors did not conduct analytics to show why so. To answer/confirm this questions/suspicions, in-depth analytics by human experts are needed to look at both false positive and false negative cases. Such an effort would also generate better 'gold standard' on this cohort for transgender phenotyping algorithms.

Unfortunately, we do not have access to chart review in COAST for further investigation of false positive and false negative cases. As described in the Discussion, we would expect to see relatively high specificity (few false positives) and low sensitivity (many false negatives) in any transgender status case finding algorithm as there are no clinical applications for assigning a transgender specific diagnosis to a non-transgender individual (the primary cause of a false positive). We highlight this primary issue in the Discussion, as well as other factors related to provider billing and coding systems, and point to next steps in this line of research as gender affirming care is increasingly integrated into Family Medicine practice, likely improving sensitivity of transgender case finding algorithms in the future. In light of reviewer comments, we have also included a discussion of the lower PPV for the best-performing CP compared to the CP based on recent transgender diagnoses; suggesting the DTP provider based transgender measure has better coverage for recent cases, and the potential benefit of recent diagnoses over ever in future CP development (page 9-10).

The authors did mention a related work of using natural language processing (NLP) in transgender phenotype computations. However, as significant amount of data is hidden in free-text records, this work would benefit from more discussions on this topic [1-3]: (1) whether better signals could be surfaced from free text to support better gold standard; (2) whether/how NLP and machine learning can be combined with their coded data to achieve better performances.

We have now included a discussion of the issues raised by the reviewer on page 10, integrating some of the literature suggested below.

It would be beneficial to link the work to the UK-wide reusable phenotype effort led by Health Data Research UK: https://portal.caliberresearch.org/.

1.Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e231. doi:10.1136/amiajnl-2013-001926

2.Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory: Table 1. Journal of the American Medical Informatics Association. 2013;20(e2):e226-e231. doi:10.1136/amiajnl-2013-001926

11.Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015;350(apr24 11):h1885-h1885. doi:10.1136/bmj.h1885

3. Wu H, Hodgson K, Dyson S, et al. Efficient Reuse of Natural Language Processing Models for Phenotype-Mention Identification in Free-text Electronic Medical Records: A Phenotype Embedding Approach. JMIR Medical Informatics 2019;7:e14782. doi:10.2196/14782

Reviewer: 2

Reviewer Name: Mei-Sing Ong
Institution and Country: Harvard Medical School & Harvard Pilgrim Health Care Institute
Please state any competing interests or state 'None declared': None declared.

Ashleigh et al. developed computable phenotypes for identifying transgender patients from EHR data. This is an important contribution, given the lack of research on this population. The derived computable phenotypes achieved high specificity but poor sensitivity, which dampens my enthusiasm for the manuscript. There may be opportunities to enhance the algorithms by incorporating other diagnoses and there are also some methodological issues that need be address (please see comments below).

Thank you, we appreciate the thorough review. In light of reviewer comments we have included additional discussion of the findings and implications of high specificity and low sensitivity of CPs in this study in the Discussion (page 9). We contend the CP developed in this feasibility study had good success in identifying cases for transgender samples, albeit with room for improvement in terms of sensitivity, and as such represents an important advancement of opportunities for transgender health research- in particular investigation of transgender health relative to other groups. Please also see response to Reviewer 1 above with regards to the high specificity but poor sensitivity findings.

Introduction:

- The paragraph under "Methodological limitations in transgender health research" is unrelated to the subject under study.

We have revised the Introduction to clarify that, in light of multiple methodological limitations in transgender health research, CP methods such as those used in the current study represent a unique opportunity to overcome some of these limitations to produce large, cost-effective transgender samples for health research with rich healthcare utilization data.

- Page 4: line 49 – the statement "computable phenotypes (CPs), also called natural language processing algorithms or case algorithms" is incorrect. Computable phenotype and natural language processing (NLP) are not synonymous. NLP is a technique for text mining, which can be applied to develop computable phenotypes. This current study did not use NLP. "Case ascertainment algorithms" may be a more appropriate term than "case algorithms".

- Page 4: line 51 – the statement "a computable phenotype is a clinical feature … that can be determined directly from EHR…" is not accurate. Rather, a computable phenotype is an algorithm for identifying clinical features using EHR data.

We have revised the above mentioned two sections of the manuscript as recommended by the reviewer.

Methods:

- Since physician-assigned transgender status (i.e. the gold standard used for model validation) was available only for HIV patients, how did the authors validate the transgender status of non-HIV patients?

We were only able to validate the CP for the HIV-positive patients using the provider-based transgender status . We have revised the manuscript throughout to ensure that this is clear to readers.

- What was the mean follow-up time of each patient? The study did not appear to impose a minimum duration of follow-up in the selection criteria. Duration of follow-up can affect whether or not there was sufficient data for case ascertainment.

We did not limit eligibility to a minimum follow-up as this was an exploratory feasibility study with an expected small sample size. In future studies with larger EMR samples, we plan to investigate the impact of follow-up time on case ascertainment.

- Page 6 (lines 39-42): “To assess face validity and utility of diagnosis and prescription data over time in CP development, concordance analyses evaluated the presence of at least one included diagnosis and prescription during the COAST study follow-up period with the presence of at least one included diagnosis and prescription in the last study year.” It’s unclear what was being evaluated here? Did the authors intend to evaluate whether or not diagnosis and prescription data were consistent over time? For what purpose? Please clarify.

We have revised this section (page 6) to further clarify that the purpose of the concordance analysis was to assess the face validity of the CP, specifically whether the identified transgender sample had exogenous sex hormone prescription use and other diagnosis patterns consistent with that of the transgender population in the existing literature.

- The authors stated that concordance analysis was performed, but did not actually quantify the degree of concordance. The Kappa statistic can be a useful measure here.

We calculated simple concordance, specifically the number of individuals who were concordant over the total number of individuals assessed. This information was used to assess the face validity of the algorithm in identifying transgender individuals with hormone prescription use and other diagnoses typical of the transgender population in the literature. The concordance analysis was not intended to otherwise validate the use of hormone prescriptions or other diagnoses as measures to identify transgender individuals. As such, calculating additional concordance statistics such as Kappa was outside the scope of the concordance analysis in this study.

- Page 7 (lines 18-20): “to further assess face validity of the transgender CP for future health research, descriptive statistics were calculated …” It was not clear to me why this step was done until I read the study results.
Please clarify here that a comparison was made between the study cohort vs the general population.

We have revised this section (page 7) to further clarify that we characterized the demographic and chronic condition profile of the total transgender sample (HIV-positive and HIV-negative).

- The authors mentioned that androgen blocker and sex hormone can be prescribed for the treatment of other conditions (e.g. hypertension, menopausal symptoms). Have the authors considered including these variables into their computable phenotypes as a way of improving the performance?

We did indeed include androgen blockers and exogenous sex hormones in the tested CPs, as described in the Methods section under Measures & Analyses, Transgender computable phenotypes sub-section (page 6). The full list of androgen blockers and sex hormones included in the CPs is detailed in the supplementary material as well.

Results:

- In Table 1, "Unspecified endocrine disorder use" should be "unspecified endocrine disorder".
- Page 8, line 4: should reference to Table 1 be Table 2 instead?

Thank you, we have made these revisions as suggested on pages 7 and 8.

Discussion:

- Page 9, lines 32-37: "Though a relatively small proportion of the "true" transgender sample was identified in this study, the impact on future analyses comparing health outcomes for transgender and cisgender group is likely negligible…" I would disagree with this statement. The current algorithm missed a large proportion of transgender patients and individuals who received less healthcare services were probably more likely to be left out. And so studies that apply the derived computable phenotype to identify transgender patients may not be able to detect these disparities.

We have clarified this section (page 9) to more precisely describe the meaning of the "true" transgender sample to be as classified by the gold standard validation measure. We agree that the CP in this study, as with all other transgender diagnosis-based CPs, is unable to capture the full transgender population, particularly those transgender people not accessing medical and/or surgical transition. We discuss this issue further in the paper, particularly in the Limitations section (page 11).

VERSION 2 – REVIEW

REVIEWER	Honghan Wu University College London, UK
REVIEW RETURNED	15-Nov-2020

GENERAL COMMENTS	All my comments have been addressed reasonably. I am happy with the edits.

REVIEWER	Mei-Sing Ong Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute, United States
REVIEW RETURNED	30-Nov-2020

GENERAL COMMENTS	Thank you for clarifying the questions raised in the previous review. The revised manuscript is much clearer. However, several important weaknesses of the study have not been addressed: 1. The performance of the computable phenotypes is really suboptimal. More importantly, the cohort is unlikely to be

	representative of transgender population at large. The performance of the algorithms did not change when diagnosis codes and hormone therapy were used as selection criteria, vs diagnosis codes alone, suggesting that the computable phenotypes were capable of identifying only individuals who received hormone therapy. This has serious implications on the generalizability of the proposed algorithms, since many transgender individuals do not receive hormone therapy and untreated patients may have different healthcare needs than those who were treated. Given this potential bias, the argument that “at worst misclassification will bias future analyses toward the null” is difficult to justify. The implications of these limitations should at the very least be acknowledged. 2. There may be ways to improve the sensitivity of the computable phenotypes – e.g. inclusion of patients who received hormone therapy in the absence of other conditions that may be treated with the same therapy, with or without the appropriate diagnosis codes; inclusion of other gender-affirming therapies. These can be easily explored. A more in-depth analysis of individuals who were missed by the computable phenotypes (as suggested by Reviewer 1) could also shine light on how the algorithms may be improved. This can be done even without access to medical charts – e.g. by comparing diagnosis codes and procedures. It may be that structured data alone cannot adequately identify the cohort of interest – this should be acknowledged and discussed. 3. It is still unclear how the authors identified transgender individuals in the HIV-negative cohort, given that provider-reported transgender measure was available only in the HIV-positive cohort. Although the revised manuscript now specified that CP validation was applied only on the HIV-positive cohort, it appears that the HIV-negative cohort was still used in the analysis and reported in the results. 4. The analysis did not impose a follow-up period and was therefore subject to biases. Furthermore, not knowing the follow-up period limits the reproducibility of the analysis, hence the utility of the computable phenotypes. This is a serious limitation that can be easily addressed, or at the very least acknowledged and discussed. The mean and range of follow-up period should also be reported.

VERSION 2 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1

Reviewer Name: Honghan Wu

Institution and Country: University College London, UK

Comments to the Author:

All my comments have been addressed reasonably. I am happy with the edits.

Thank you, we appreciate the review.

Reviewer: 2

Reviewer Name: Mei-Sing Ong

Institution and Country: Department of Population Medicine, Harvard Medical School & Harvard Pilgrim Health Care Institute, United States

Comments to the Author:

Thank you for clarifying the questions raised in the previous review. The revised manuscript is much clearer.

However, several important weaknesses of the study have not been addressed:

1. The performance of the computable phenotypes is really suboptimal. More importantly, the cohort is unlikely to be representative of transgender population at large. The performance of the algorithms did not change when diagnosis codes and hormone therapy were used as selection criteria, vs diagnosis codes alone, suggesting that the computable phenotypes were capable of identifying only individuals who received hormone therapy. This has serious implications on the generalizability of the proposed algorithms, since many transgender individuals do not receive hormone therapy and untreated patients may have different healthcare needs than those who were treated. Given this potential bias, the argument that “at worst misclassification will bias future analyses toward the null” is difficult to justify. The implications of these limitations should at the very least be acknowledged.

Thank you for the review. In response to the reviewer’s feedback, we have revised the sentence cited above by the reviewer “at worst misclassification would bias results related to disparities between transgender and cisgender health toward the null….” (page 9) to further emphasize that the results of the current study apply

in the context of transgender health research studies using administrative data (and relying on diagnostic criteria ascertainment measures). We would also like to highlight that ‘true’ in this context

refers to those transgender people identified via the ‘gold standard’ DTP provider based transgender status measure in the validation analysis, not in fact all transgender people in BC (page 9).

We agree with the reviewer, the transgender population identified in this feasibility study using administrative health data is unlikely to be representative of the transgender population in its entirety, but instead the transgender sub-population accessing transition related care (approximately 24-47% of the total transgender population per Scheim & Bauer 2015 J Sex Res, as referenced in the manuscript). Transgender case identifying algorithms like the one produced in this current study are largely only capable of identifying those accessing transition related healthcare services as they rely on diagnostic criteria for case ascertainment. We include a section discussing this limitation in the manuscript Discussion (page 11). As we note in the manuscript, Collin and colleagues published an important study (2016 J Sex Med) reviewing the impact of study design and ascertainment measure on transgender status measurement in the transgender health literature. Their study found the prevalence of transgender status to vary widely depending on study design and ascertainment measure (discussed on page 10 of our manuscript), and that studies employing diagnostic criteria for transgender status ascertainment resulted in an undercount of transgender populations compared to other studies (discussed on page 11 of our manuscript). Ultimately, study type and ascertainment measure are a function of available data as well as research questions and study objectives. While results from our study are likely not representative of the total transgender population, they are likely generalizable to other studies investigating transgender health research questions with administrative data-the appropriate application context for case-finding algorithms.

Administrative data is an invaluable resource in health research, enabling large samples of rich clinical data needed to answer essential questions about the health and health services use of populations. While the use of administrative data for investigating transgender health research questions has its limitations, it also offers important benefits - including the potential for capture of large sample sizes impossible with observational research with this small ‘hidden’ population. We contend that while CP performance was sub-optimal overall, this study successfully meets the objective to demonstrate the feasibility of identifying transgender people in administrative data in Canada for the first time and points the way forward for future studies applying similar methods.

2. There may be ways to improve the sensitivity of the computable phenotypes – e.g., inclusion of patients who received hormone therapy in the absence of other conditions that may be treated with the same therapy, with or without the appropriate diagnosis codes; inclusion of other gender-affirming therapies. These can be easily explored. A more in-depth analysis of individuals who were missed by the computable phenotypes (as suggested by Reviewer 1) could also shine light on how the algorithms may be improved. This can be done even without access to medical charts – e.g., by comparing diagnosis codes and procedures. It may be that structured data alone cannot adequately identify the cohort of interest – this should be acknowledged and discussed.

We agree that structured data alone is insufficient to identify all transgender people in an administrative dataset. However, the current study was unable to incorporate free text from health records or charts for the case finding algorithms as only structured EMR data is linked through COAST (as described on page 10 of the Discussion). In response to the reviewer’s feedback, we have added a discussion of this limitation of structured data in the current study, but the potential

opportunities for NLP and machine learning approaches using data sources that include both structured data and free text EMR data in the Discussion, “CP development and validation” section (page 10).

3. It is still unclear how the authors identified transgender individuals in the HIV-negative cohort, given that provider-reported transgender measure was available only in the HIV-positive cohort. Although the revised manuscript now specified that CP validation was applied only on the HIV-positive cohort, it appears that the HIV-negative cohort was still used in the analysis and reported in the results.

As the reviewer states, this study included both the COAST HIV-negative and HIV-positive cohorts for the identification of transgender individuals. However, the validation analysis of the algorithm for the identification of transgender individuals was only conducted using the COAST HIV-positive cohort where there was an available independent measure of transgender status via the Drug Treatment Program (DTP) provider-based measure. As the DTP is the provincial HIV case registry and comprises patients who have been diagnosed with HIV and accessed treatment in British Columbia, the validation analysis was feasible with the COAST HIV-positive cohort only. The best performing CP from the validation analysis was then applied to both COAST cohorts to identify a transgender sample of mixed HIV-serostatus. These methodological details have been updated in the Methods section to further clarify (page 7).

4. The analysis did not impose a follow-up period and was therefore subject to biases. Furthermore, not knowing the follow-up period limits the reproducibility of the analysis, hence the utility of the computable phenotypes. This is a serious limitation that can be easily addressed, or at the very least acknowledged and discussed. The mean and range of follow-up period should also be reported.

As suggested by the reviewer, we have now reported mean and range follow-up time for each of the CPs tested in the validation analysis in the manuscript (Results section, Validation sub-section, page 8). We have also revised the manuscript to reflect this inclusion (Methods section, Validation sub-section, page 7). As reported, follow-up time was similar across the four CPs tested, with the second longest mean follow-up time corresponding to the best-performing CP. While differences in study follow-up time can introduce bias, impact study validity and reliability, and affect case-finding algorithm performance, the similar mean and range duration of follow-up time for all CPs assessed suggests that differential follow-up time was not a source of significant bias in this study. We have included discussion of these elements in the Discussion, CP development and validation sub-section, page 10.